Real Time Adaptive Event Detection in Astronomical Data Streams: Lessons from the VLBA

David R. Thompson[1,2]
Sarah Burke-Spolaor[2]
Adam T. Deller[3]
Walid A. Majid[2]
Divya Palaniswamy[4]
Steven J. Tingay[4]
Kiri L. Wagstaff[1]
Randall B. Wayth[4]

[1] david.r.thompson@jpl.nasa.gov
[2] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA
[3] ASTRON, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
[4] International Centre for Radio Astronomy Research, Curtin University. GPO Box U1987, Perth WA, 6845. Australia

A new generation of observational science instruments is dramatically increasing collected data volumes in a range of fields. These instruments include the Square Kilometre Array (SKA), Large Synoptic Survey Telescope (LSST), terrestrial sensor networks, and NASA satellites participating in "decadal survey" missions. Their unprecedented coverage and sensitivity will likely reveal wholly new categories of unexpected and transient events. *Commensal* methods passively analyze these data streams, recognizing anomalous events of scientific interest and reacting in real time. We report on a case example: V-FASTR, an ongoing commensal experiment at the Very Long Baseline Array (VLBA) that uses online adaptive pattern recognition to search for anomalous *fast radio transients*. V-FASTR triages a millisecond-resolution stream of data and promotes candidate anomalies for further offline analysis. It tunes detection parameters in real time, injecting synthetic events to continually retrain itself for optimum performance. This self-tuning approach retains sensitivity to weak signals while adapting to changing instrument configurations and noise conditions. The system has operated since July 2011, making it the longest-running real time commensal radio transient experiment to date.

1. Introduction

The next generation of scientific instruments will dramatically increase collected data volumes in multiple disciplines. Astronomers will soon face petascale data streams from instruments like the Large Synoptic Survey Telescope and Square Kilometer Array [Dewdney 2009]. The environmental remote sensing community will see orders of magnitude increases in data volume from "decadal survey" satellite missions such as OCO-2, HyspIRI and GeoCAPE. These instruments will almost certainly observe secondary transient events and anomalies of scientific interest alongside their primary missions. Examples of transient activities include orbiters that serendipitously observe wildfires or erupting volcanoes and ground-based telescopes that observe supernovae and other momentary astronomical events. These generally require immediate action, either to trigger additional observations or to archive contextual data needed for validation. The time window for followup action is typically quite short, demanding efficient real time detection algorithms.

This work describes a case study investigating *fast radio transients*. These millisecond-scale pulses of radio frequency energy may be produced by astronomical events such as Gamma Ray Bursts, intermittent radio pulsars known as RRATs, or possibly more extreme events such as merging black holes [Cordes2004, Lazio2009]. At the time of this writing, few candidate transients have been discovered in past archives [Keane2012], and there have been no wholly conclusive detections of the most exotic sources. More examples are needed to validate these detections and

characterize the population. To this end, passive commensal analyses, which operate in piggyback on other investigations, are an appealing way to accumulate significant telescope time.

Radio transient detection typifies common challenges of large-scale commensal searches. One is consistency. Commensal analysis must cope with a range of different interference conditions and configurations. It must also keep pace with any instrument upgrades and modifications. Another challenge is data volume. The system operates continuously with high duty cycle and processing loads. This requires an automated system that can triage the data stream, promoting only the most important segments to storage and analyst attention.

Here we describe the Very Long Baseline Array Fast Transients Experiment (V-FASTR), a system for commensal detection of fast radio transient events [Wayth2011, Thompson2011]. V-FASTR is a pathfinder project for a larger transient detection system known as CRAFT. It analyzes data from the Very Long Baseline Array (VLBA) to search for transient events and anomalies. Previous work analyzed its detection algorithms using simulation and controlled tests prior to deployment [Thompson2011]. Here we report on the actual performance during the first year of operations. We also describe a novel new approach for *self-tuning* detection parameters by injecting synthetic events into the data stream. This adapts to a range of instrument conditions, detecting rare, subtle astronomical events and improving resilience to noise. As of this writing, V-FASTR is the single longest running commensal transient search to date and an instructive example of real time online event detection.

2. The V-FASTR system

The Very Long Baseline Array (VLBA) consists of ten identical antennas at locations across the western hemisphere [Romney2010]. Typical VLBA observations point all of the telescopes at a common location to perform interferometric imaging with very high spatial resolution. Each antenna records raw voltage samples that are later sent in physical disks to the array operations center in Socorro, NM. Here they are time-aligned, *correlated*, and processed to form an interferometric radio sky image. After this step the disks are erased for reuse.
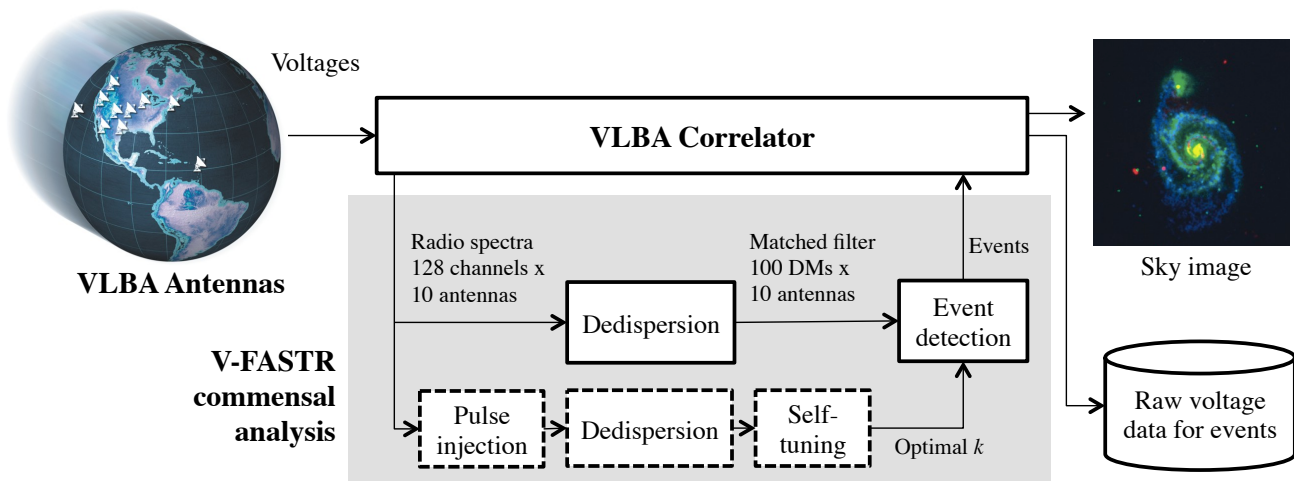
Figure 1: V-FASTR design and interfaces with the VLBA Correlator. The dashed boxes are the adaptive self-tuning system, which runs in alongside the main detection pipeline. Images courtesy of NRAO/AUI.

V-FASTR must identify any transients in the short interval before the disks are erased. The original raw voltage data is important for validation because it retains the phase information of the signal, allowing for vital post-analyses such as imaging (finding the physical position of a pulse on the sky), coherent summing (combining phased signals to improve sensitivity), and coherent dedispersion (removing dispersive smoothing in the voltage domain). Unfortunately, voltage data is so voluminous that it is impossible to archive more than a small fraction. The V-FASTR commensal detection system [Wayth2011] solves this by searching for events in real time and saving only the most promising data. It acts as a plugin to the DiFX software correlator [Deller2007], which we have modified to broadcast

antenna autocorrelations over a secondary network. In this mode the correlator serves as a spectrometer measuring the channelized, squared, and accumulated power at each antenna across 32 frequency channels. The integration time interval is approximately 1 ms. V-FASTR searches this data stream and notifies the correlator of any detected events. At the end of the correlation process the system saves a copy of these candidate segments to long-term storage for more thorough analysis. Figure 1 illustrates the V-FASTR system design. The commensal analysis runs on a rack-mounted computer with six modern processors but typically uses only a fraction of these resources. We describe each software component in greater detail below.

2.1 Dedispersion Search The first stage of the V-FASTR analysis is a *dedispersion search*, effectively a matched filter search based on knowledge of how signals propagate through the interstellar medium. It corrects for the distortion caused by free electrons in the space between the source and the receiver. Following [Lyne1998], a broadband EM pulse experiences a frequency-dependent delay:

$$(1) \qquad t_2 - t_1 = 4.15\text{ms DM}[(\nu_1/\text{GHz})^{-2}) - (\nu_2/\text{GHz})^{-2})]$$

The delay in observed signal arrival between two frequency channels ν_1 and ν_2 is dependent on the bandwidth, the center frequency, and the Dispersion Measure (DM) of the source. This last value is related to distance and the density of the interstellar medium. The DM is not known for a new source. Consequently, the V-FASTR system performs a *dedispersion search* over many candidate DMs [Battacharya1998]. At each timestep it tests a range of possible dispersion measures, equivalent to applying a bank of matched filters with appropriate frequency/time profiles. Dedispersion searching has proven effective for the detection of pulsars and other astronomical phenomena.

A dedispersion search can also help separate genuine astronomical signals from Radio Frequency Interference (RFI). Local interference typically manifests as a broadband signal with no dispersion, since the pulse originates locally and all frequencies arrive simultaneously. Figure 2 (Left) is typical of the input data arriving to the V-FASTR system from the correlator. This dynamic spectrum image displays the data as a matrix of signal powers channelized into discrete time and frequency bins. The pixel intensity shows observed power, the accumulated squared voltage measured by a station. This segment contains several kinds of local interference such as narrow-band RFI limited to a few frequency channels and a bright momentary broadband signal near 0.14 seconds. The panel at right shows a real astrophysical event: a single pulse from the pulsar J1919+0021, which has an actual DM of 90.315. Note that the dispersion bends the pulse slightly; the lower frequency components experience greater dispersion delay. Searching over many DMs transforms the initial spectral data into a vector of approximately 100 matched filter scores - one per DM searched - for each of 10 antennas.
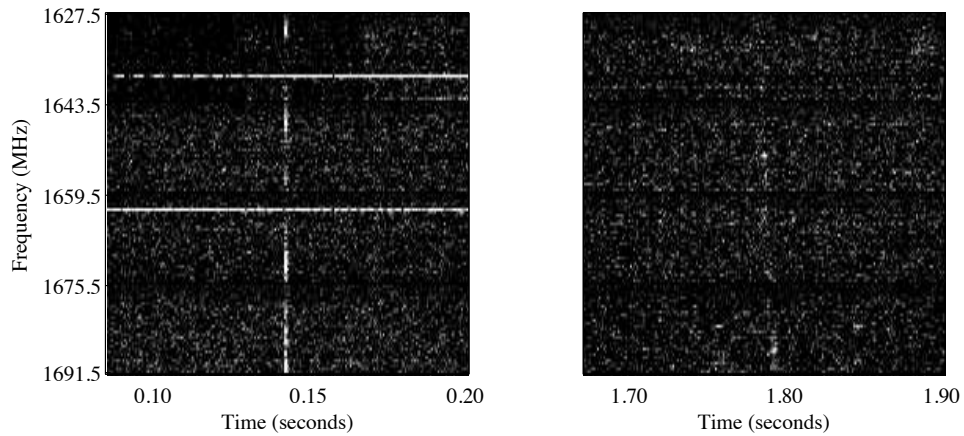
Figure 2: Examples of radio interference, including channelized, momentary, and band-level changes. Right: A single pulse from pulsar J1919+0021, with a classic dispersed pulse profile. This plot uses an incoherent sum across all VLBA stations to improve contrast; in reality the dedispersion search analyzes each station's data stream separately. The times shown here are consistent with the dedispersed time series shown later in Figure 3.

2.2 Robust detection. The VLBA was designed to have low system temperatures and good gain stability over long timescales. It is conventionally used in a low time resolution regime. In contrast, V-FASTR measures signals at high time resolution and experiences additional interference due to momentary RFI, background signal discontinuities, varying gains, and other artifacts. We must account for these effects in order to limit the number of false detections.

We exploit the fact that the stations are geographically separated, so interference only appears at a single station while true astrophysical signals are correlated across many stations [Thompson2011]. The top nine rows of Figure 3 show the signal from each antenna after dedispersion to a DM of 90. An arrow shows a real astronomical pulse from the pulsar J1919+0021. The pulse appears in multiple stations simultaneously, indicating that it is isolated on the celestial sphere. Because we do not know the correct DM in advance, many tens of these time series are produced simultaneously and the system must analyze them all. This provides a real-valued input vector having dimensionality of order one hundred. The task amounts to a multivariate pattern recognition problem with millisecond time requirements, a nonstationary background distribution, and a very low tolerance for false positives (10^{-5} false positives per timestep).
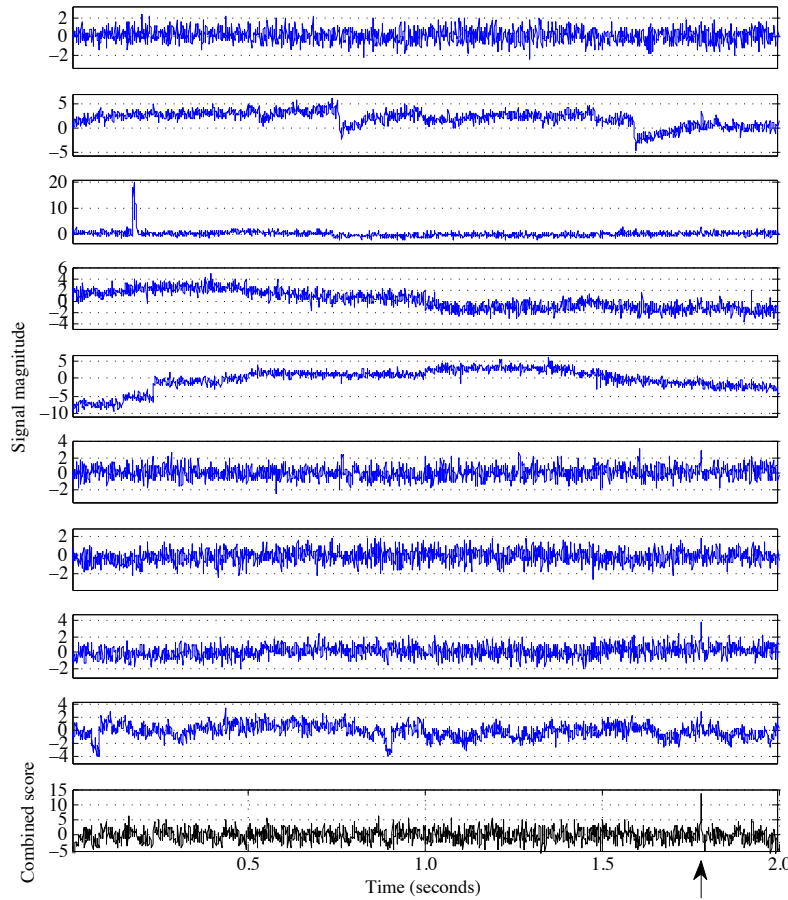
Figure 3: Detection of a single pulse from pulsar J1919+0021. The top 9 panels show dedispersed time series from different VLBA antennas. Non-astrophysical artifacts and interference events appear at several stations. Such differences are often caused by RFI but could also be caused by maintenance issues, weather, and the source position above each station's horizon. The bottom panel shows the scalar output of the adaptive detection score. An arrow indicates the event - the only apparent pulse from this source. It was missed by the standard detection approach but successfully found by the V-FASTR system.

We represent each sample in the dedispersed time series as a multi-element vector $\mathbf{x} \in \mathbb{R}^{d \times n}$ where d is the number of dispersion measures and n the number of stations. We can formalize the detection rule as a function $f_k(\mathbf{x})$ that maps

the dedispersed time series to a scalar score which is high for true transients but low for interference. We anticipate that any astrophysical fast transients will be very rare. For instance [Madau1998, Siemion2011] estimate the rate in a local 1 Gpc^3 volume to be one per square degree of sky per 1000 hours. For the VLBA primary beam field of view of 0.27 deg^2 and 1 ms temporal quantization this amounts to a target/background fraction of approximately 1 x 10^{-10} [Wayth2011, Wayth2012].

The conventional approach to signal detection in VLBA data is to sum the signals from all antennas. As described in previous work [Thompson2011], the V-FASTR system uses a simple *robust estimator* that excises one or more extreme values and averages the rest. First the data stream from each station is noise-whitened, standardized to have zero mean and unit standard deviation, dedispersed, and finally filtered using a high-pass filter with a 100 ms diameter. We then apply a separate matched filter for each DM and store the result as a multivariate time series. Suppose $\mathbf{x}_{id}$ is the matched filter result from station i at dispersion measure d. We sort the antennas in order of increasing signal strength for each DM and sum all but the most extreme k station signals to make a combined signal. The maximum response over all DMs becomes the overall score for that timestep:

$$(2) \qquad \text{Detect if } f_k(\mathbf{x}) \geq \tau \quad \text{for} \quad f_k(\mathbf{x}) = \sup_d \frac{1}{(n-k)} \sum_{i=1}^{n-k} \mathbf{x}_{id}$$

This is most appropriate if the stations provide independent and identically distributed estimates of some common source perturbed by occasional sporadic noise. Operationally we define τ in terms of a Signal to Noise Ratio (SNR) given in standard deviations from the mean.

The best value of k depends on noise statistics. V-FASTR uses incoherent addition that ignores phase information, so the measured SNR for a real event approximately follows the relationship $\sqrt{n-k}$ and excising too many antennas reduces sensitivity. However, excising too few extreme values increases the false positive rate. The optimal tradeoff depends on background noise and the number of antennas participating in the observation.

We formalize the expected loss for a particular value of k using Bayesian decision theory. Each timestep is associated with a background event y_0 or a transient event y_1, with transient events much rarer than the background. We posit a misclassification cost c_0 for a false positive detection, and c_1 for missing a real transient event. We optimize k and τ to minimize Bayes' loss, i.e. the expected loss considering both error probabilities.

$$(3) \qquad L(\tau, k) \quad = \quad c_1 \, p(f_k(\mathbf{x}) < \tau \,|y_1) \, p(y_1) + c_0 \, p(f_k(\mathbf{x}) \geq \tau \,|y_0) \, p(y_0)$$

$$(4) \qquad\qquad\qquad = \quad \alpha \, p(f_k(\mathbf{x}) < \tau \,|y_1) + p(f_k(\mathbf{x}) \geq \tau \,|y_0)$$

The α term is a constant that incorporates the prior probability of each event as well as the relative costs of misclassification. Even positing a very large weight for c_1, the class imbalance is much larger and we can assume $\alpha \ll 1$.

2.3 Adaptive self-tuning for RFI excision. It is challenging to construct a universal detection rule because Equation 3 depends on noise properties that are unknown in advance and difficult to model. Instead we use the *data stream itself* for an empirical nonparametric estimate of the loss function $L(\tau,k)$. Specifically, we use Equation 1 to construct synthetic pulses at known locations and dispersions, add these signals to the input stream and then attempt to recover them using a range of parameterizations. We set τ to be the minimum value that does not produce any significant false positives (e.g., very low-α regimes). We use the number of synthetic pulses recovered prior to the first m false positives, for low m, as a figure of merit to select k.

During operations the commensal system buffers 10,000 timesteps of spectrometer data from the correlator and retrains itself on the preceding batch prior to detection. It injects 200 dispersed pulses at SNRs from 5 to 9 and DMs from 10 to 50. This range does not span the full dedispersion search, but it allows the system to fit more pulses into the same time interval. The parallel stream is then dedispersed and analyzed with all possible excision levels. We compute the detection figure of merit permitting a small number m of real positives (typically 20) so that single, isolated real pulses do not significantly affect the result. Typically the system excises a single antenna but can remove up to four in scans having very high interference. In Figure 3 the system elected to excise two antennas, resulting in the combined detection score shown in the bottom panel.
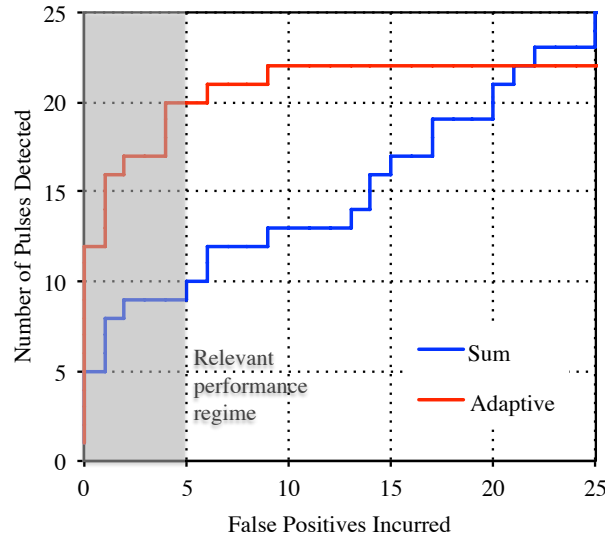
3. Performance

This section evaluates the adaptive system's performance. The VLBA occasionally observes known pulsars during its normal operations, which provides a blind sensitivity test. To date, the commensal system has detected hundreds of events from seven different pulsars (Table 1) In contrast, the conventional non-adaptive approach of summing signals has failed to detect the two weakest pulsars (J1919+0021 and J0146+5922). Figure 3 shows the J1919+0021 detection.

Pulsar Name	Maximum SNR	DM	On-pulse flux density (Jy)
J0332+5434	>50	26	14.0
J1136+1551	20	5	1.0
J0826+2637	19	20	0.5
J1935+1616	12	158	1.0
J01645-0317	10	36	0.5
J0157+6212	10	30	0.14
J1919+0021	9	90	0.05
J0147+5922	9	40	0.13
J0953+0755	8	3	2.0

Table 1: Known pulsars detected serendipitously by V-FASTR.

We also compared detection results for observations containing many pulses. Figure 4 shows a Receiver Operating Characteristic (ROC) curve for a scan containing over 20 pulses from J0826+2637. Part of the curve is irrelevant to our task since the correlator can only archive a small number of events in the time interval between detection and disk erasure. Typically it saves at most four or five events with a priority ordering determined by their signal strength. Consequently the settings with fewer than five false positives, shaded in gray, are most important. The adaptive approach finds a far greater number of real astrophysical events meaning it successfully prioritizes weak real pulses over RFI. We have tuned the adaptive system for maximum precision in the critical low-false-positive regime.


Figure 4: Detection performance for pulsar J0826+2637 by adaptive and non-adaptive (summing) approaches.

A major bottleneck to sustained commensal analyses is the limited analyst time available to review any candidate detections. The V-FASTR policy is to manually review all detections with a signal-to-noise ratio above 7. *Post-hoc* analyses of V-FASTR performance show that the adaptive system has greatly reduced the overall number of these candidates that must be reviewed. April 2012 provided a good test case since there were no pulsars or other known

astrophysical transient events being observed. We applied a standard RFI filter to remove all signals with a DM less than 10. Any remaining detections in April are false positives. Figure 5 shows the number of detections at different SNR thresholds, for both adaptive and non-adaptive approaches. The non-adaptive approach yields an unfeasibly large number of detections for the same SNR threshold.
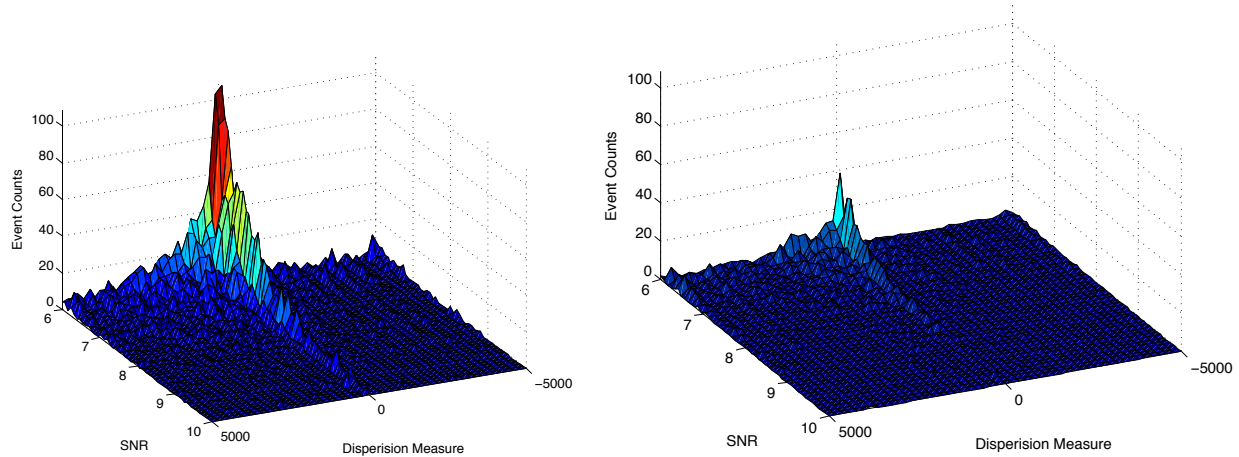
Figure 5: The total count of false positive detections for the month of April 2012. The vertical axis shows the number of detections while the horizontal axes show SNR and DM values. Left: False positives detected by the standard summing approach. Right: False positives detected by the V-FASTR adaptive method.

Finally, we recorded the excision levels actually selected by the self tuning system during this period. Several very clean observations at high frequencies resulted in the system choosing k = 0, tantamount to no RFI protection. These scans did not produce any false positive events. For detected events, the proportions of excision levels in use were 54.7% (k = 1), 23.9% (k = 2), 10.6% (k = 3), and 10.7% (k = 4). Over 78% of the detections used an excision level of just one or two stations. The larger excision levels appear only during extremely noisy observations or those containing pulsed signals from satellites. Satellite signals can appear in multiple stations simultaneously, and are currently the largest source of false positives.

4. Discussion, Conclusions, and Lessons Learned

V-FASTR performs adaptive detection of fast transients by commensal analysis at the Very Long Baseline Array. Its ability to retrain automatically through injection of synthetic events has improved reliability over traditional detection methods. This system has operated for over one year and has recently placed new limits on the rates of astrophysical phenomena observed at 1.4 GHz and above [Wayth2012]. This makes it a working example of adaptive pattern recognition in data streams to enable new science results. V-FASTR continues to accumulate telescope time at multiple frequencies; manageable false positive rates mean that the system is sustainable for the long term.

The effective composition and coordination of the team has been crucial factor in this successful deployment. The V-FASTR group combines experts in machine learning with experts in radio astronomy and the VLBA's core processing system. This fusion of expertise kept the machine learning approaches grounded in the application and operational perspectives. As a result, the VLBA has adopted V-FASTR for full-time use, establishing the policy that all observing campaigns have their data analyzed by the commensal system by default. Naturally, any interesting detections made by V-FASTR would immediately result in the involvement of that campaign's Principal Investigator. The provision of immediate data access to the automated system is a notable policy innovation that paves the way for more commensal automated science concepts. Other instruments and telescopes have expressed interest in similar systems to aid in their own processing pipelines.

5. Acknowledgements

References

[Bhattacharya1998] D. Bhattacharya. NATO ASIC Proc. 515: The Many Faces of Neutron Stars, 1998.

[Cordes2007] J. Cordes, T. Lazio, and M. McLaughlin. New Astronomy Reviews, 48(11):1459–1472, 2004. A. Deller, S. Tingay, M. Bailes, and C. West. Difx: a software correlator for very long baseline interferometry using multiprocessor computing environments. Publications of the Astronomical Society of the Pacific, 119:318–336, 2007.

[Dewdney2012] P. Dewdney, P. Hall, R. Schilizzi, and T. Lazio. Proceedings of the IEEE, 97(8):1482– 1496, 2009.

[Keane2012] E. F. Keane, B. W. Stappers, M. Kramer, and A. G. Lyne. ArXiv e-prints, 1206.4135, June 2012.

[Lazio2009] J. Lazio, J. Bloom, G. Bower, J. Cordes, S. Croft, S. Hyman, C. Law, and M. McLaughlin. The dynamic radio sky: An opportunity for discovery. In Astro2010: The Astronomy and Astrophysics Decadal Survey, page 176. 2009.

[Lyne1998] A. G. Lyne and F. G. Graham-Smith. Pulsar Astronomy. Cambridge University Press, 1998.

[Madau1998] P. Madau, M. Della Valle, and N. Panagia. On the evolution of the cosmic supernova rates. Monthly Notices of the Royal Astronomical Society, 297(1):L17–L22, 1998.

[Romney2010] J. D. Romney, editor. Very Long Baseline Array Observational Status Summary. NRAO, 2010. http://www.vlba.nrao.edu/astro/obstatus/current/obssum.html.

[Siemion2011] A. Siemion, G. Bower, G. Foster, P. McMahon, M. Wagner, D. Werthimer, D. Backer, J. Cordes, and J. van Leeuwen. The allen telescope array fly's eye survey for fast radio transients. The Astrophysical Journal, 744(2):109, 2011.

[Thompson2011] D. R. Thompson, K. L. Wagstaff, W. F. Brisken, A. T. Deller, W. A. Majid, S. J. Tingay, and R. B. Wayth. The Astrophysical Journal, 735:98, 2011.

[Wayth2011] R. B. Wayth, W. Brisken, A. Deller, W. Majid, D. Thompson, S. Tingay, and K. Wagstaff. The Astrophysical Journal, 735:97, 2011.

[Wayth2012] R. B. Wayth, S. J. Tingay, A. T. Deller, W. F. Brisken, D. R. Thompson, K. L. Wagstaff, and W. A. Majid. The Astrophysical Journal (in press), Arxiv preprint arXiv:1205.5840, 2012.

Biographies

Dr. David R. Thompson is a research technologist at the Jet Propulsion Laboratory, California Institute of Technology and a member of IEEE. His research applies machine learning to remote exploration.

Dr. Sarah Burke Spolaor is a post-doctoral astrophysics researcher at the Jet Propulsion Laboratory, California Institute of Technology. Her research focuses on pulsar astronomy and detecting short-timescale radio transients.

Dr. Adam T. Deller is a VENI postdoctoral fellow at ASTRON, the Netherlands Institute for Radio Astronomy. His research interests include high precision astrometry of compact radio sources and instrumentation development for Very Long Baseline Interferometry.

Dr. Walid A. Majid is a senior researcher at the Jet Propulsion Laboratory, California Institute of Technology. His research involves pulsar astronomy and development of instrumentation and signal processing techniques for radio astronomy.

Divya Palaniswamy is a PhD candidate at Curtin University whose thesis work focuses on high time resolution radio astronomy.

Dr. Steven J. Tingay is a Professor of Radio Astronomy at Curtin University in Perth, Western Australia, and a Western Australian Premier's Research Fellow. His research focuses on radio astronomy instrumentation and astrophysics.

Dr. Kiri L. Wagstaff is a senior researcher at the Jet Propulsion Laboratory, California Institute of Technology, where she focuses on the use of machine learning methods to investigate science questions using planetary and astronomical data.

Dr. Randall B.Wayth is a research scientist at Curtin University in Western Australia, and a member of IEEE. His research interests include radio astronomy instrumentation and astrophysics.

Author contact information

David R. Thompson – david.r.thompson@jpl.nasa.gov - Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

Sarah Burke-Spolaor - Sarah.Burke-Spolaor@jpl.nasa.gov - Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

Adam T. Deller - deller@astron.nl - ASTRON, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands

Walid A. Majid - walid.a.majid@jpl.nasa.gov - Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

Divya Palaniswamy - divya.palaniswamy@postgrad.curtin.edu.au - International Centre for Radio Astronomy Research, Curtin University. GPO Box U1987, Perth WA, 6845. Australia

Steven J. Tingay - steven.tingay@icrar.org - International Centre for Radio Astronomy Research, Curtin University. GPO Box U1987, Perth WA, 6845. Australia

Kiri L. Wagstaff - kiri.l.wagstaff@jpl.nasa.gov - Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

Randall B. Wayth - r.wayth@curtin.edu.au - International Centre for Radio Astronomy Research, Curtin University. GPO Box U1987, Perth WA, 6845. Australia